Practical use of aggregator data quality metrics in a collection scenario

Andrew Bentley ‡

‡ University of Kansas, Lawrence, KS, United States of America

Abstract

The recent incorporation of standardized data quality metrics into the GBIF, iDigBio, and ALA portal infrastructures enables data providers with useful information they can use to clean or augment Darwin Core data at the source based on these recommendations. Numerous taxonomic and geographic based metrics provide useful information on the quality of various Darwin Core fields in this realm, while also providing input on Darwin Core compliance for others. As a provider/data manager for the Biodiversity Institute, University of Kansas, and having spent some time evaluating their efficacy and reliability, this presentation will highlight some of the positive and negative aspects of my experience with specific examples while highlighting concerns regarding the user experience and standardization of these metrics across the aggregator landscape. These metrics have indicated both data and publishing issues that have increased the utility and cleanliness of our data while also highlighting batch processing challenges and issues with the process of inferring "bad" data. The integration of these metrics into source database infrastructure will also be postulated, with Specify Software as an example.

Keywords

Aggregators, GBIF, iDigBio, metrics, data quality, collections, IPT

Presenting author

Bentley, Andrew C

Presented at

SPNHC/TDWG 2018

Acknowledgements

University of Kansas Biodiversity Institute (KUBI)

Specify Software Project (http://www.sustain.specifysoftware.org/)

Funding program

DBI - ADVANCES IN BIO INFORMATICS

Grant title

ABI Sustaining: Supporting Biological Collections Computing withSpecify